

# The Convergence Rate of Majority Vote under Exchangeability

Miles E. Lopes

mlopes@stat.berkeley.edu

Department of Statistics

University of California, Berkeley

## Abstract

Majority vote plays a fundamental role in many applications of statistics, such as ensemble classifiers, crowdsourcing, and elections. When using majority vote as a prediction rule, it is of basic interest to ask “How many votes are needed to obtain a reliable prediction?” In the context of binary classification with Random Forests or Bagging, we give a precise answer: If  $\text{err}_t$  denotes the test error achieved by the majority vote of  $t \geq 1$  classifiers, and  $\text{err}^*$  denotes its nominal limiting value, then under basic regularity conditions,  $\text{err}_t = \text{err}^* + \frac{c}{t} + o(\frac{1}{t})$ , where  $c$  is a constant given by a simple formula. More generally, we show that if  $V_1, V_2, \dots$  is an exchangeable Bernoulli sequence with mixture distribution  $F$ , and the majority vote is written as  $M_t = \text{median}(V_1, \dots, V_t)$ , then  $1 - \mathbb{E}[M_t] = F(\frac{1}{2}) + \frac{1}{8}F''(\frac{1}{2})\frac{1}{t} + o(\frac{1}{t})$  when  $F$  is sufficiently smooth.

## 1 Introduction

Majority vote is a core principle for aggregating decisions. At an abstract level, votes are a statistical resource, which may be obtained for a cost, such as computation, communication, or time. As more votes are collected, the majority vote is typically more likely to select the “correct” candidate, but at a higher cost. This trade-off leads to the basic statistical problem of determining the smallest number votes needed to make a reliable decision.

An important instance of this general problem arises in the context of *ensemble methods* for binary classification. Well-known examples of ensemble methods include Bagging, Boosting, and Random Forests (Breiman, 1996, 2001; Freund and Schapire, 1995). The connection between voting and ensemble methods arises in the following way. Given a fixed set of  $N_0$  labeled training examples  $\mathcal{D} = \{(X_j, Y_j)\}_{j=1}^{N_0}$  in a sample space  $\mathcal{X} \times \{0, 1\}$ , an algorithm is used to train an ensemble of  $t \geq 1$  *base classifiers*  $Q_i : \mathcal{X} \rightarrow \{0, 1\}$ ,  $i = 1, \dots, t$ , which are often randomized. The predictions of the base classifiers are then aggregated by a particular rule, with majority vote being the standard choice for Bagging and Random Forests. More concretely, if a test point  $(X, Y)$  is sampled from  $\mathcal{X} \times \{0, 1\}$  with  $Y$  being unknown, then the prediction of the whole ensemble is given by the median of the predicted labels  $Q_1(X), \dots, Q_t(X)$ . We denote the test error by  $\text{err}_t = \mathbb{P}(\text{median}(Q_1(X), \dots, Q_t(X)) \neq Y \mid \mathcal{D})$ , and always assume that  $t$  is odd to eliminate the issue of ties.

For many ensemble methods, the test error  $\text{err}_t$  typically decreases and then stabilizes as the number of base classifiers becomes large ( $t \rightarrow \infty$ ). Likewise, the nominal limiting value, say  $\text{err}^*$ , is viewed as a target level of performance. As  $\text{err}_t$  approaches  $\text{err}^*$ , we also pay an increasing computational cost — since each base classifier must be separately trained, stored in memory, and evaluated for new predictions. Furthermore, the cost is often compounded by the need to carry out the entire procedure for a variety of different tuning parameters. Consequently, it is natural to select the smallest number  $t^*$  such that  $|\text{err}_{t^*} - \text{err}^*|$  is less than a given tolerance, which amounts to determining the convergence rate of  $\text{err}_t$ . This is the problem we aim to solve in the present paper, with particular emphasis on the methods of Bagging and Random Forests.

Despite the close connection between the convergence rate of  $\text{err}_t$  and the computational cost of an ensemble, this issue has received little attention in the literature for Bagging and Random Forests. Regarding the more distinct method of Boosting, a substantial amount of work has been done to analyze its rate of convergence (Bickel et al., 2006; Mukherjee et al., 2011; Schapire, 2010; Zhang and Yu, 2005). However, the aggregation rule used in Boosting is very different from ordinary majority vote, since the Boosting algorithm iteratively reweights the votes as the ensemble is grown. Our work here focuses on the majority voting rule when reweighting does not occur, and our results are not comparable to convergence rates for Boosting.

The most closely related work that we are aware of is a paper by Ng and Jordan (2001), which analyzes the convergence rate of the so-called “voting Gibbs classifier” — a Bayesian ensemble method that generates labels  $Q_1(X), \dots, Q_t(X)$  from a posterior distribution, and then aggregates them via majority vote. The error measure studied by Ng and Jordan is essentially a Bayesian version of  $\text{err}_t$ , and they prove that its convergence rate is at most  $\mathcal{O}(\frac{1}{t})$  under basic regularity conditions. With regard to the problem of choosing  $t$ , their analysis techniques do not seem to be generally applicable, since they do not specify a constant or the exact rate of convergence. Apart from that paper, we are not aware of similar results for other ensemble methods involving majority vote.

In this paper, our primary contribution is a formula for  $\text{err}_t$  that is *exact* to order  $\frac{1}{t}$ . The formula is applicable to any ensemble method based on the majority vote of an exchangeable sequence of labels (cf. Assumption 1 below). In fact, the statement of our result in Theorem 1 extends beyond the context of classification, and may be relevant to aggregation problems in other areas, such as recommender systems, online markets, or social choice theory (Easley and Kleinberg, 2010). Given that many voting models are analyzed under the restrictive assumption of i.i.d. votes, our much weaker assumption of exchangeability also lends itself to applications. Further discussion of exchangeable voting models in social choice theory may be found in Berg (1993); Ladha (1993); Zaigraev and Kaniovski (2012).

The statement of our main result is given with proof in the following section. Technical lemmas are proved in Section 3.

## 2 Main results

To define the test error in precise terms, there are three sources of randomness to consider: the training set  $\mathcal{D}$ , the randomized base classifiers  $Q_i(\cdot)$ , and the test point  $(X, Y)$ . For the purposes of our analysis, the randomness in  $\mathcal{D}$  will play no role, and all of our probability statements will be conditional on  $\mathcal{D}$ . Even though  $\mathcal{D}$  is viewed as fixed, the functions  $Q_i(\cdot)$  may depend on additional randomization in the training algorithm. For instance, the Bagging and Random Forests algorithms draw a random subset of  $\mathcal{D}$  to train each  $Q_i(\cdot)$ . Lastly, the test point  $(X, Y)$  is sampled as a pair from  $\mathcal{X} \times \{0, 1\}$ , independently of  $\mathcal{D}$  and the functions  $Q_i(\cdot)$ . Altogether, if we let  $Q = (Q_1, \dots, Q_t)$  and write the joint distribution of  $(X, Y, Q)$  as  $\mathbb{P}_{(X, Y, Q)}$ , then we define

$$\text{err}_t := \mathbb{P}_{(X, Y, Q)}(\text{median}(Q_1(X), \dots, Q_t(X)) \neq Y \mid \mathcal{D}). \quad (1)$$

The subscript on  $\mathbb{P}_{(X, Y, Q)}$  will be omitted from now on.

The main technical challenge of studying  $\text{err}_t$  arises from the correlation structure of the labels  $Q_1(X), Q_2(X), \dots$ , which may be very complex in general. Nevertheless, in the cases of Random Forests and Bagging, the correlation structure is constrained by the fact that the labels form an *exchangeable Bernoulli sequence* (cf. de Finetti's Theorem (Billingsley, 2012, Theorem 35.10)). In particular, Random Forests and Bagging obey the following assumption.

**Assumption 1.** *The sequence of binary labels  $Q_1(X), Q_2(X), \dots$  is conditionally i.i.d., given the test point  $(X, Y)$  and the training data  $\mathcal{D}$ .*

In fact, this condition can be found in implicit form in the seminal Random Forests paper of (Breiman, 2001, Definition 1.1), and it has been used elsewhere as an abstract definition of Random Forests (Biau et al., 2008).

By restricting our attention to the class of ensemble methods that satisfy Assumption 1, our analysis of  $\text{err}_t$  can be reduced to the study of exchangeable Bernoulli sequences in the following way. First, if we define  $\pi_0 := \mathbb{P}(Y = 0)$  and  $\pi_1 := \mathbb{P}(Y = 1)$  as the class proportions, then  $\text{err}_t$  may be decomposed as a weighted sum of class-wise error rates

$$\begin{aligned} \text{err}_t = & \pi_0 \mathbb{P}(\text{median}(Q_1(X), \dots, Q_t(X)) = 1 \mid \mathcal{D}, Y = 0) \\ & + \pi_1 \mathbb{P}(\text{median}(Q_1(X), \dots, Q_t(X)) = 0 \mid \mathcal{D}, Y = 1). \end{aligned} \quad (2)$$

Next, we define  $\{U_i\}$ ,  $i = 1, 2, \dots$  to be the sequence of predicted labels for a random test point drawn from the negative class,

$$\{U_i\} \stackrel{\mathcal{L}}{=} \{Q_i(X)\} \mid \{\mathcal{D}, Y = 0\}. \quad (3)$$

Similarly, we define  $\{\tilde{U}_i\}$  to be the sequence of predicted labels for a random test point drawn from the positive class,

$$\{\tilde{U}_i\} \stackrel{\mathcal{L}}{=} \{Q_i(X)\} \mid \{\mathcal{D}, Y = 1\}. \quad (4)$$

It is clear from Assumption 1 that the sequences  $\{U_i\}$  and  $\{\tilde{U}_i\}$  are exchangeable. When  $t$  is odd, we may also write

$$\text{err}_t = \pi_0 \mathbb{E}[\text{median}(U_1, \dots, U_n)] + \pi_1 (1 - \mathbb{E}[\text{median}(\tilde{U}_1, \dots, \tilde{U}_n)]). \quad (5)$$

Of course, there is no formal distinction between the two terms on the right hand side (apart from labeling). Therefore, it is natural to state our main result in terms of the more basic object at hand: the running median of a generic exchangeable Bernoulli sequence. This also serves to emphasize that our result is broadly applicable to other voting scenarios, and is not limited to ensemble methods.

We now fix some remaining notation for the statement of Theorem 1. Recall from de Finetti's theorem that if  $V_1, V_2, \dots$  is an exchangeable Bernoulli sequence, then there exists a random variable  $\Theta$  taking values in the interval  $[0, 1]$ , such that the variables  $V_1, V_2, \dots$  are conditionally i.i.d. Bernoulli( $\theta$ ), given  $\Theta = \theta$ . If we let  $F : [0, 1] \rightarrow [0, 1]$  denote the distribution function of the variable  $\Theta$ , then we refer to  $F$  as the *mixture distribution* for the sequence  $\{V_i\}$ .

**Theorem 1.** *Let  $V_1, V_2, \dots$  be an exchangeable Bernoulli sequence with mixture distribution  $F$ , and let  $M_t = \text{median}(V_1, \dots, V_t)$ . Suppose  $F$  is twice continuously differentiable on  $[0, 1]$ . Then as  $t \rightarrow \infty$ ,*

$$1 - \mathbb{E}[M_t] = F(\tfrac{1}{2}) + \tfrac{1}{8}F''(\tfrac{1}{2})\tfrac{1}{t} + o(\tfrac{1}{t}). \quad (6)$$

In order to extract the convergence rate of  $\text{err}_t$  from the theorem, a few more pieces of notation are needed. We denote the mixture distributions of the predicted label sequences  $\{U_i\}$  and  $\{\tilde{U}_i\}$  by  $G$  and  $\tilde{G}$  (respectively), and we identify them with their distribution functions from  $[0, 1]$  to  $[0, 1]$ . We also define the constants

$$c := \tfrac{\pi_1}{8}\tilde{G}''(\tfrac{1}{2}) - \tfrac{\pi_0}{8}G''(\tfrac{1}{2}), \quad (7)$$

$$\text{err}^* := \pi_0(1 - G(\tfrac{1}{2})) + \pi_1\tilde{G}(\tfrac{1}{2}). \quad (8)$$

Due to the relation (5), our formula for  $\text{err}_t$  is obtained directly from Theorem 1. Note also that the convergence  $\text{err}_t \rightarrow \text{err}^*$  follows from our result, and is not an assumption.

**Corollary 1.** *Suppose  $G$  and  $\tilde{G}$  are twice continuously differentiable on  $[0, 1]$ . Then as  $t \rightarrow \infty$ ,*

$$\text{err}_n = \text{err}^* + \tfrac{c}{t} + o(\tfrac{1}{t}). \quad (9)$$

In light of the corollary, it is natural to wonder what meaning the functions  $G$  and  $\tilde{G}$  have in terms of the classification problem. The idea may be explained in the following way. If  $x \in \mathcal{X}$  is any fixed point in the sample space, then we define the function  $\vartheta : \mathcal{X} \rightarrow [0, 1]$  by

$$\vartheta(x) := \mathbb{E}[Q_1(x) \mid \mathcal{D}]. \quad (10)$$

Hence, if  $x$  is a point that should be labeled with “1”, then  $\vartheta(x)$  represents the average accuracy of the ensemble at that point. (Note that the variables

$Q_1(x), Q_2(x), \dots$  are i.i.d.) Having defined  $\vartheta$ , it is easy to see that  $G$  and  $\tilde{G}$  are the distribution functions of the following random variables,

$$G \stackrel{\mathcal{L}}{=} \vartheta(X) \mid \{\mathcal{D}, Y = 0\}, \quad (11)$$

$$\tilde{G} \stackrel{\mathcal{L}}{=} \vartheta(X) \mid \{\mathcal{D}, Y = 1\}. \quad (12)$$

Viewing  $\vartheta$  as an “accuracy function” on  $\mathcal{X}$ , our differentiability assumptions on  $G$  and  $\tilde{G}$  express the idea that there is a smooth transition between regions of  $\mathcal{X}$  that are easy to classify and regions that are difficult to classify. Specifically, the existence of  $G''(1/2)$  and  $\tilde{G}''(1/2)$  implies that the test points  $(X, Y)$  assign their mass smoothly over the boundary of “ambiguous points” where  $\vartheta(x) = 1/2$ . Note too that this smoothness condition depends on both the ensemble algorithm, and on the way the test points  $(X, Y)$  are distributed over  $\mathcal{X} \times \{0, 1\}$ . In this sense, the functions  $G$  and  $\tilde{G}$  offer a very compact way of encoding all in the information in the problem that is relevant to  $\text{err}_t$ .

We conclude this section by turning to the proof of Theorem 1. The main technique involved is to represent  $\mathbb{E}[M_t]$  in terms of a second order Edgeworth expansion for the binomial distribution function. Although it is also possible to study  $\mathbb{E}[M_t]$  using a simple Hoeffding bound, that approach seems to lead to an inferior rate of  $\mathcal{O}(\frac{1}{\sqrt{t}})$ .

*Proof of Theorem 1.* We begin by writing

$$\begin{aligned} 1 - \mathbb{E}[M_t] &= \mathbb{P}\left(\frac{1}{t} \sum_{i=1}^t V_i \leq 1/2\right) \\ &= \int_0^1 \mathbb{P}\left(\frac{1}{t} \sum_{i=1}^t V_i \leq 1/2 \mid \Theta = \theta\right) dF(\theta), \end{aligned} \quad (13)$$

where we note that the integrand is a binomial distribution function. Our aim is to evaluate the limit of the scaled difference

$$t(1 - \mathbb{E}[M_t] - F(1/2)) = \int_0^1 t(\mathbb{P}(\frac{1}{t} \sum_{i=1}^t V_i \leq 1/2 \mid \Theta = \theta) - F(1/2)) dF(\theta), \quad (14)$$

and show that the limit is equal to  $\frac{1}{8}F''(\frac{1}{2})$ .

The first main portion of the proof involves reducing the last integral to a more concrete form. If we let  $\mathcal{E}_t$  denote the second order Edgeworth expansion for the distribution function of  $\text{Binomial}(t, \theta)$ , then Lemma 1 in Section 3 guarantees the following uniform approximation for any  $\theta \in (0, 1)$ ,

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{t}}{\sqrt{\theta(1-\theta)}}\left(\frac{1}{t} \sum_{i=1}^t V_i - \theta\right) \leq z \mid \Theta = \theta\right) - \mathcal{E}_t(z) \right| = o(t^{-1}). \quad (15)$$

The boundary cases of  $\theta = 0, 1$  will play no role, since the continuity of  $F$  implies that  $\Theta$  hits the endpoints of  $[0, 1]$  with zero probability. Using the uniformity of the condition (15), and letting

$$z = z(\theta; t) = \frac{\sqrt{t(1/2-\theta)}}{\sqrt{\theta(1-\theta)}}, \quad (16)$$

we may replace the probability  $\mathbb{P}(\frac{1}{t} \sum_{i=1}^t V_i \leq 1/2 \mid \theta)$  in line (14) with  $\mathcal{E}_t(z(\theta; t))$ . Hence,

$$t(1 - \mathbb{E}[M_t] - F(1/2)) = \int_0^1 t(\mathcal{E}_t(z(\theta; t)) - F(1/2))dF(\theta) + o(1). \quad (17)$$

We also let  $\varphi_t(z)$  denote the first and second order terms of the expansion,

$$\mathcal{E}_t(z) = \Phi(z) + \varphi_t(z), \quad (18)$$

which allows us to express the scaled difference  $t(1 - \mathbb{E}[M_t] - F(1/2))$  as

$$\int_0^1 t(\Phi(z(\theta; t)) - F(1/2))dF(\theta) + \int_0^1 t\varphi_t(z(\theta; t))dF(\theta) + o(1). \quad (19)$$

Surprisingly, the higher order terms  $\varphi_t(z)$  give no contribution to the term  $\frac{1}{8}F''(\frac{1}{2})\frac{1}{t}$  of the main formula (6). In particular, Lemma 3 in Section 3 shows that the second integral  $\int_0^1 t\varphi_t(z(\theta; t))dF(\theta)$  tends to 0 as  $t \rightarrow \infty$ . Consequently, it remains to consider the limit of  $\int_0^1 t(\Phi(z(\theta; t)) - F(1/2))dF(\theta)$ .

In order to simplify the first integral in line (19), it is an essential element of the proof to notice that  $z(\theta; t)$  is a smooth and monotone function of  $\theta \in (0, 1)$ , which may be inverted for any fixed  $t$ , giving

$$\theta = \theta(z; t) = \frac{1}{2} - \frac{z/2}{\sqrt{z^2 + t}}.$$

Changing variables from  $\theta$  to  $z$ , and then integrating by parts, it follows that

$$\int_0^1 t(\Phi(z(\theta; t)) - F(1/2))dF(\theta) = \int_{-\infty}^{\infty} t(F(\theta(z; t)) - F(1/2))\phi(z)dz, \quad (20)$$

where  $\phi$  denotes the standard normal density. We note that an extra minus sign has been introduced because  $z$  is a decreasing function of  $\theta$ . Since  $\theta(0; t)$  is equal to  $1/2$  for all  $t$ , a first order Taylor expansion at  $z = 0$  gives

$$F(\theta(z; t)) - F(1/2) = F'(1/2)\theta'(0; t)z + R(z; t), \quad (21)$$

with  $R(z; t)$  denoting the remainder. Using  $\int z\phi(z)dz = 0$ , the first order term in line (21) vanishes upon integration for every  $t$ , giving

$$\int_0^1 t(\Phi(z(\theta; t)) - F(1/2))dF(\theta) = \int_{-\infty}^{\infty} tR(z; t)\phi(z)dz. \quad (22)$$

We conclude the proof by determining the pointwise limit of  $R(z; t)$  and applying the dominated convergence theorem. Writing the remainder  $R(z; t)$  in Lagrange form, we have

$$tR(z; t) = \frac{1}{2}t \cdot \left( F''(\theta(\xi; t))[\theta'(\xi; t)]^2 + F'(\theta(\xi; t))\theta''(\xi; t) \right) \cdot z^2, \quad (23)$$

for some  $\xi$  between 0 and  $z$ . Using the continuity of  $F''$  at  $1/2$ , as well as the formulas

$$\theta'(z; t) = -\frac{t/2}{(t+z^2)^{3/2}} \quad (24)$$

and

$$\theta''(z; t) = \frac{3tz}{2(t+z^2)^{5/2}}, \quad (25)$$

we obtain the following pointwise limit for each fixed  $z$  as  $t \rightarrow \infty$ ,

$$t R(z; t) \rightarrow \frac{1}{8} F''(\frac{1}{2}) z^2. \quad (26)$$

It is straightforward to check that  $t R(z; t)$  is dominated by a fixed polynomial in  $z$ , which is clearly integrable with respect to  $\phi$ . The details are given in Lemma 4 of Section 3. Consequently, the desired limit

$$\int_{-\infty}^{\infty} t R(z; t) \phi(z) dz \rightarrow \frac{1}{8} F''(\frac{1}{2}).$$

follows from the formula  $\int z^2 \phi(z) dz = 1$ .  $\square$

### 3 Technical lemmas

The technical lemmas supporting Theorem 1 are based largely on the second order Edgeworth expansion of the binomial distribution function. Since the binomial distribution arises from a sum of *lattice* variables, the expansion includes a number of terms that are absent from the usual expansion for continuous variables. The following result has been adapted from a general theorem in (Bhattacharya and Rao, 1986) (Theorem 23.1), which handles multivariate lattice distributions. Additional work is needed to obtain explicit formulas for the second order terms in their expansion, but we omit these tedious calculations. Formulas for the second order terms have also been given in (Brown et al., 2002) (Lemma 1), with the only difference being that we have written the expansion in terms of Hermite and Bernoulli polynomials to simplify the proof of Lemma 2 below.

To state the first lemma, several pieces of notation are needed. We define the parameter

$$\rho_t = \rho_t(\theta, z) := \llbracket \frac{1}{2} - t\theta + \sqrt{t}\sigma z \rrbracket, \quad (27)$$

where  $\sigma^2 := \theta(1 - \theta)$  and the symbol  $\llbracket \alpha \rrbracket$  denotes the fractional part of a real number  $\alpha$ . The first two Bernoulli polynomials are

$$\begin{aligned} B_1(\rho_t) &= \rho_t - \frac{1}{2} \\ B_2(\rho_t) &= \rho_t^2 - \rho_t + \frac{1}{6}, \end{aligned} \quad (28)$$

and the first several Hermite polynomials are

$$\begin{aligned} H_1(z) &= z \\ H_2(z) &= z^2 - 1 \\ H_3(z) &= z^3 - 3z \\ H_4(z) &= z^4 - 6z^2 + 3 \\ H_5(z) &= z^5 - 10z^3 + 15z. \end{aligned} \quad (29)$$

Lastly, if  $\kappa_i$  denotes the  $i$ th cumulant of the Bernoulli( $\theta$ ) distribution, then the “standardized cumulants” are given by

$$\frac{\kappa_3}{\sigma^3} = \frac{1-2\theta}{\sqrt{\theta(1-\theta)}} \quad \text{and} \quad \frac{\kappa_4}{\sigma^4} = \frac{6\theta^2-6\theta+1}{\theta(1-\theta)}. \quad (30)$$

**Lemma 1** (Bhattacharya and Rao, 1976). *Let  $W_1, W_2, \dots$ , be i.i.d. Bernoulli( $\theta$ ) variables with  $\theta \in (0, 1)$ . For any  $z \in \mathbb{R}$  and  $t \geq 1$ , define*

$$G_t(z) := \mathbb{P} \left( \frac{\sqrt{t}}{\sqrt{\theta(1-\theta)}} \left( \frac{1}{t} \sum_{i=1}^t W_i - \theta \right) \leq z \right). \quad (31)$$

Then,

$$\sup_{z \in \mathbb{R}} |G_t(z) - \mathcal{E}_t(z)| = o(t^{-1}), \quad (32)$$

where  $\mathcal{E}_t$  is the second order Edgeworth expansion given by

$$\begin{aligned} \mathcal{E}_t(z) = & \Phi(z) - \phi(z) \left( \frac{1}{6\sqrt{t}} \frac{\kappa_3}{\sigma^3} H_2(z) + \frac{1}{24t} \frac{\kappa_4}{\sigma^4} H_3(z) + \frac{1}{72t} \left( \frac{\kappa_3}{\sigma^3} \right)^2 H_5(z) \right) \\ & - \phi(z) \left( \frac{1}{\sqrt{t}\sigma} B_1(\rho_t) + \frac{1}{6t} \frac{\kappa_3}{\sigma^4} H_3(z) B_1(\rho_t) + \frac{1}{t\sigma^2} \frac{1}{2} H_1(z) B_2(\rho_t) \right). \end{aligned} \quad (33)$$

**Remarks.** We note that if the terms involving  $B_1$  and  $B_2$  were omitted, then the expansion (33) would exactly match the well-known formula for the case of continuous variables, which may be found in (McCullagh and Nelder, 1989, p. 474).

The next lemma gives formulas for the higher order terms of the expansion (33) under the change of variable  $z = \frac{\sqrt{t}(1/2-\theta)}{\sqrt{\theta(1-\theta)}}$ . In particular, it is possible to write the cumulants of Bernoulli( $\theta$ ) as functions of  $z$  when this relationship is inverted. For simplicity, we have presented the formula only for odd values of  $t$ . Apart from some added technical detail, the arguments that rely on this lemma are no different in the case of even  $t$ .

**Lemma 2.** *When  $z = \frac{\sqrt{t}(1/2-\theta)}{\sqrt{\theta(1-\theta)}}$  and  $\theta \in (0, 1)$ , the cumulants of Bernoulli( $\theta$ ) satisfy*

$$\frac{\kappa_3}{\sigma^3} = \frac{2z}{\sqrt{t}} \quad \text{and} \quad \frac{\kappa_4}{\sigma^4} = \frac{4z^2}{t} - 2, \quad (34)$$

and the parameter  $\rho_t$  defined in line (27) satisfies  $\rho_t = \lfloor t/2 \rfloor$ . Also, for this choice of  $z$  and odd values of  $t$ , the function  $\varphi_t(z) = \mathcal{E}_t(z) - \Phi(z)$  is given by

$$\varphi_t(z) = \phi(z) \left( \frac{z}{3t} H_2(z) + \frac{1}{12t} \left( \frac{2z^2}{t} - 1 \right) H_3(z) + \frac{4z^2}{72t^2} H_5(z) - \left( \frac{z^2+t}{6t} \right) H_1(z) \right). \quad (35)$$

**Remarks.** The proof of this lemma only involves algebraic manipulations and is hence omitted. The main simplification that occurs for the case of odd  $t$  is that  $B_1(\rho_t) = \rho_t - \frac{1}{2}$  vanishes identically (hence removing two terms from line (33)).

### 3.1 The higher order terms vanish as $t \rightarrow \infty$ .

**Lemma 3.** *Assume the conditions of Theorem 1 hold. If  $z(\theta; t) = \frac{\sqrt{t}(1/2-\theta)}{\sqrt{\theta(1-\theta)}}$ , and  $\varphi_t(z) = \mathcal{E}_t(z) - \Phi(z)$ , then as  $t \rightarrow \infty$ ,*

$$\int_0^1 t \varphi_t(z(\theta; t)) dF(\theta) \rightarrow 0. \quad (36)$$



*Proof.* Changing variables from  $\theta$  to  $z$ , and integrating by parts gives

$$\int_0^1 t \varphi_t(z(\theta; t)) dF(\theta) = - \int_{-\infty}^{\infty} t \varphi'_t(z) F(\theta(z; t)) dz, \quad (37)$$

where it is simple to check that the boundary term vanishes for any fixed  $t$  using the formula (35). To consider the right side of line (37), it follows from the formula (35) and the relation

$$\frac{d}{dz}[\phi(z)H_j(z)] = -\phi(z)H_{j+1}(z),$$

that the  $\mathcal{O}(\frac{1}{t})$  terms of  $\varphi'_t(z)$  can be written in the form  $\frac{1}{t}\phi(z)H_i(z)H_j(z)$ ,  $i \neq j$ , (up to constants) where we note that  $H_0(z) \equiv 1$  and  $H_1(z) = z$ . Since the functions  $F(\theta(z; t))$  are uniformly bounded by 1 and converge pointwise to the constant  $F(1/2)$  as  $t \rightarrow \infty$ , the dominated convergence theorem and the orthogonality of Hermite polynomials imply that the integral  $\int t \varphi'_t(z) F(\theta(z; t)) dz$  converges to 0.  $\square$

### 3.2 The remainder is dominated.

To simplify the statement and proof of the following lemma, we write  $a_t(z) \lesssim b_t(z)$  for two sequences of non-negative functions  $a_t(x)$  and  $b_t(x)$  if there is an absolute constant  $c > 0$  such that  $a_t(x) \leq c b_t(x)$  for all  $t \geq 1$ , and all  $z \in \mathbb{R}$ .

**Lemma 4.** *Assume the conditions of Theorem 1 hold. Then, the remainder  $R(z; t)$  in line (21) satisfies the bound*

$$t|R(z; t)| \lesssim 1 + |z|^3. \quad (38)$$

*Proof.* From line (23) we have

$$t R(z; t) = \frac{1}{2} t \cdot \left( F''(\theta(\xi; t)) (\theta'(\xi; t))^2 + F'(\theta(\xi; t)) \theta''(\xi; t) \right) \cdot z^2, \quad (39)$$

where  $\xi$  is a number between 0 and  $z$ . By assumption,  $F''$  and  $F'$  are continuous functions on  $[0, 1]$ . Consequently, the quantities  $|F''(\theta(\xi; t))|$  and  $F'(\theta(\xi; t))$  are uniformly bounded by a constant. It suffices to show that  $t(\theta'(\xi; t))^2$  and  $t|\theta''(\xi; t)|$  are dominated by fixed polynomials in  $z$  for all  $\xi \in [-|z|, |z|]$  and all  $t \geq 1$ . Due to the formulas (24) and (25), we have

$$t(\theta'(\xi; t))^2 = t \left( \frac{t/2}{(t+\xi^2)^{3/2}} \right)^2 \lesssim 1, \quad (40)$$

$$t|\theta''(\xi; t)| = \frac{3t^2|\xi|}{2(t+\xi^2)^{5/2}} \lesssim |z|, \quad (41)$$

and the statement of the lemma follows easily.  $\square$

**Acknowledgements.** MEL is very grateful to Philip Kegelmeyer for many research discussions, and for stimulating interest in this problem during a practicum at Sandia National Laboratory. Peter Bickel is also thanked for research discussions. MEL gratefully acknowledges the support of the DOE CSGF Fellowship, under grant number DE-FG02-97ER25308.

## References

- S. Berg. Condorcet’s jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1):87–95, 1993.
- R.N. Bhattacharya and R.R. Rao. *Normal approximation and asymptotic expansions*, volume 64. Society for Industrial & Applied Mathematics, 1986.
- G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.
- P. J. Bickel, Y. Ritov, and A. Zakai. Some theory for generalized boosting algorithms. *The Journal of Machine Learning Research*, 7:705–732, 2006.
- P. Billingsley. *Probability and measure*, volume 939. Wiley, 2012.
- L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- L.D. Brown, T.T. Cai, and A. Dasgupta. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1):160–201, 2002.
- D. Easley and J. Kleinberg. *Networks, crowds, and markets*. Cambridge Univ Press, 2010.
- Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.
- K. K. Ladha. Condorcet’s jury theorem in light of de finetti’s theorem. *Social Choice and Welfare*, 10(1):69–85, 1993.
- P. McCullagh and J.A. Nelder. *Generalized linear models*, volume 37. Chapman & Hall/CRC, 1989.
- I. Mukherjee, C. Rudin, and R. E. Schapire. The rate of convergence of AdaBoost. arXiv:1106.6024v1 [math.OC], 2011.
- A.Y. Ng and M.I. Jordan. Convergence rates of the voting gibbs classifier, with application to bayesian feature selection. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 377–384. Morgan Kaufmann Publishers Inc., 2001.
- R. E. Schapire. The convergence rate of adaboost. In *The 23rd Conference on Learning Theory, open problem*, 2010.
- A. Zaigraev and S. Kaniovski. Bounds on the competence of a homogeneous jury. *Theory and decision*, 72(1):89–112, 2012.

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency.  
*The Annals of Statistics*, 33(4):1538–1579, 2005.